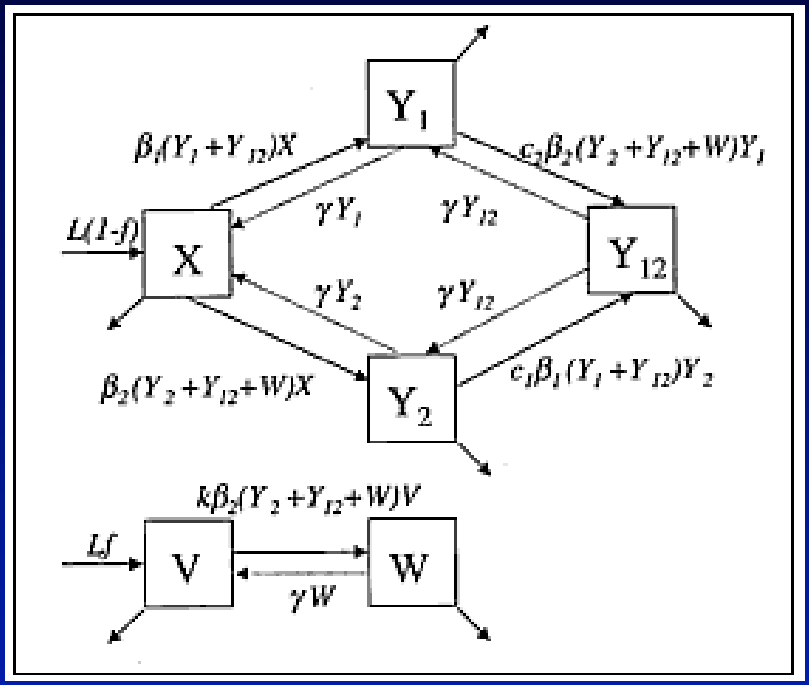
CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS

Probabilistic modeling and molecular phylogeny

Anders Gorm Pedersen

Molecular Evolution Group
Center for Biological Sequence Analysis
Technical University of Denmark (DTU)

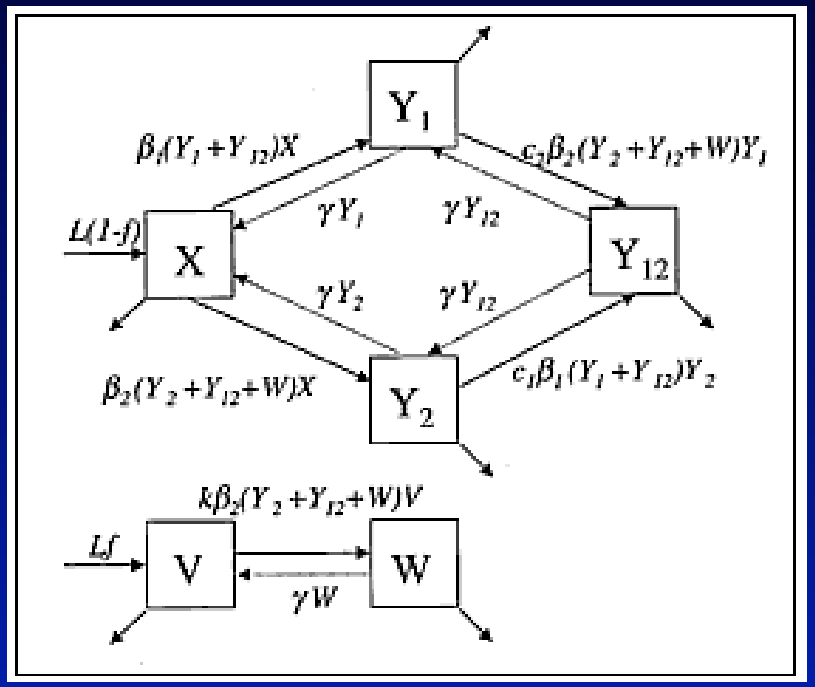
What is a model?



Mathematical models are:

- Incomprehensible
- Useless
- No fun at all

What is a model?



Mathematical models are:

- Incomprehensible
- Useless
- No fun at all

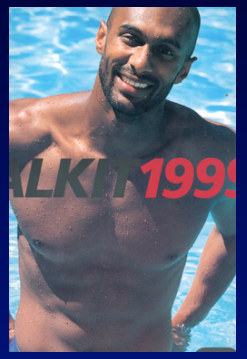
What is a Model?

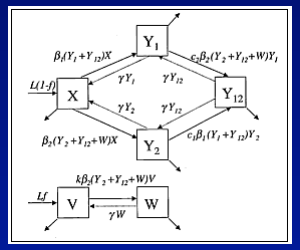
- Model = stringently phrased hypothesis !!!

- Hypothesis (as used in most biological research):
  - Precisely stated, but qualitative
  - Allows you to make qualitative predictions
  - Example: “Population size grows rapidly when there are few individuals, but growth rate declines when resources become limiting.”

- Arithmetic model:
  - Mathematically explicit (parameters)
  - Allows you to make quantitative predictions
  - Example:

$$N_t = \frac{K}{1 + \left(\frac{K}{N_0} - 1\right) e^{-rt}}$$

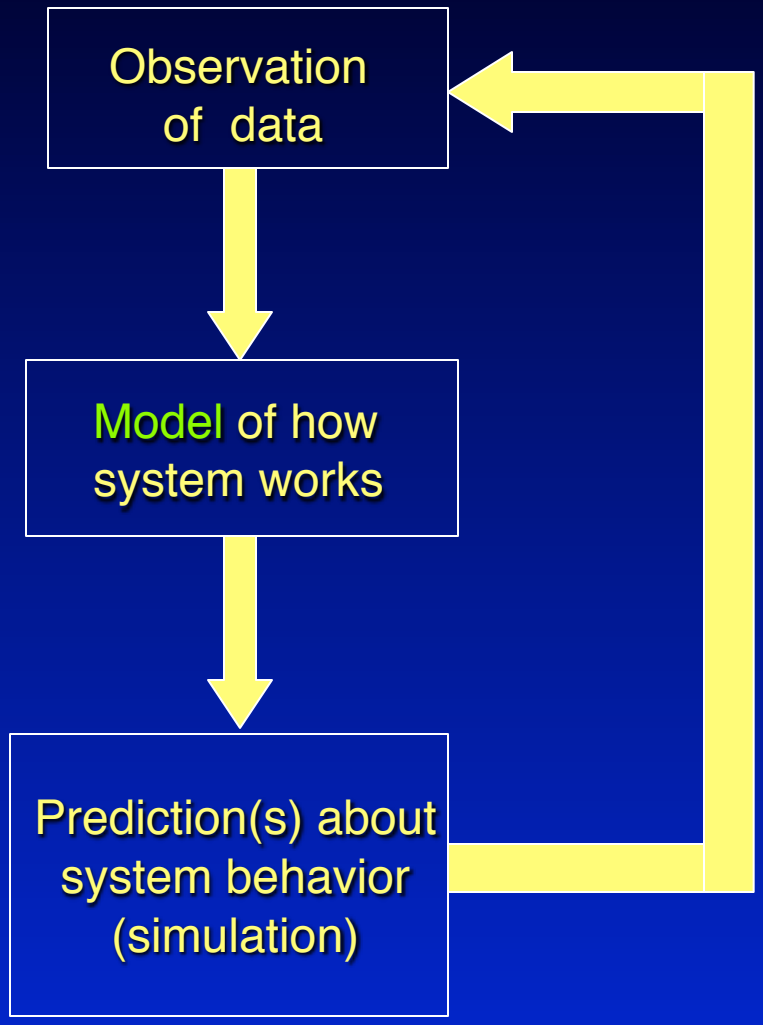




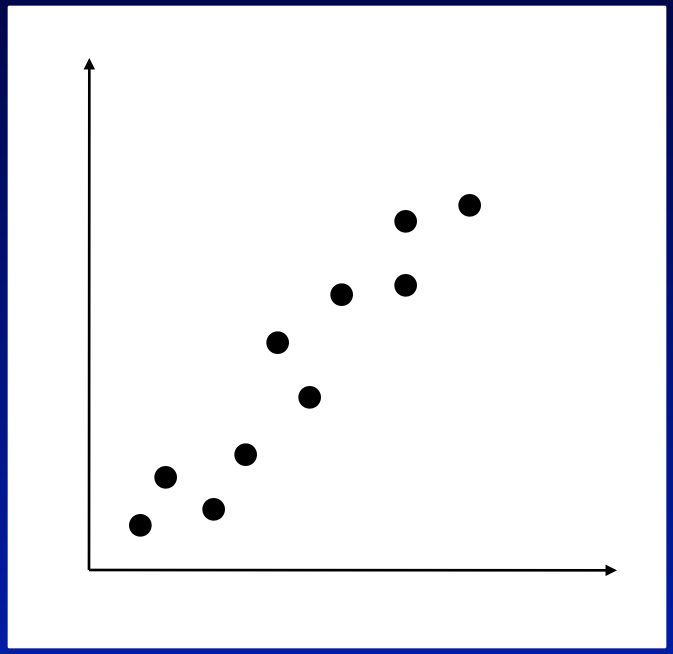
Models do not represent full reality!

- It is typically not possible to represent full reality in a mathematical model.
- Growth model example:
  - fecundity and survival rate depend on a large number of factors
  - biological and non-biological, internal and external, some stochastic
  - for each individual in a population.
  - fully realistic growth model should account for fecundity and survival rate of all individuals
  - for each individual these are complicated functions of huge numbers of different terms.
  - it is impossible to get good estimates of this multitude of parameters from a finite data set
- One-to-one maps are difficult to read!
- Goal is instead to find good approximating model
- We assume that structure of reality has factors with “tapering effect sizes”
  - a few very important factors
  - a moderate number of moderately important factors
  - very many factors of little importance

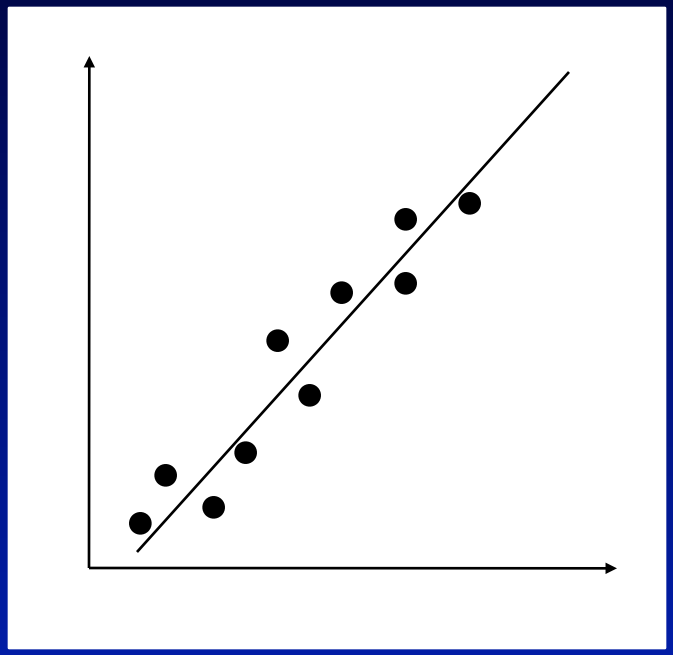
The Scientific Method



Modeling: An example



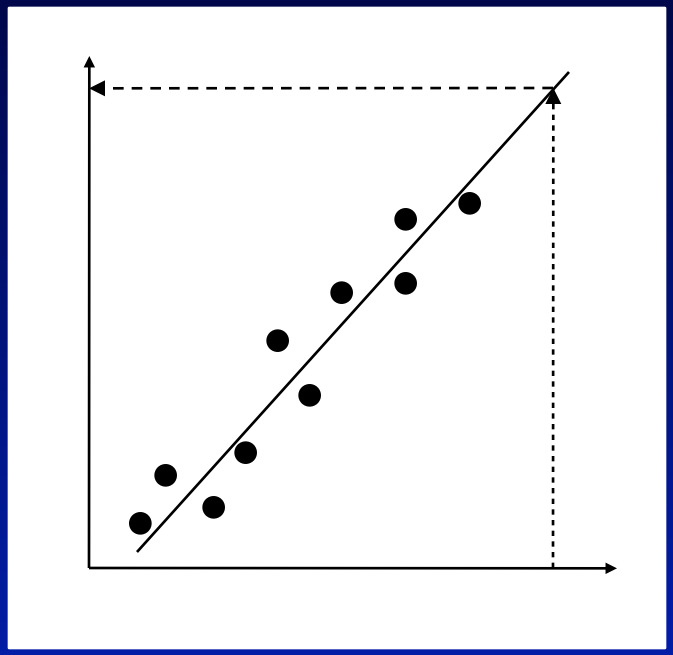
Modeling: An example



$$y = ax + b$$

Simple 2-parameter model

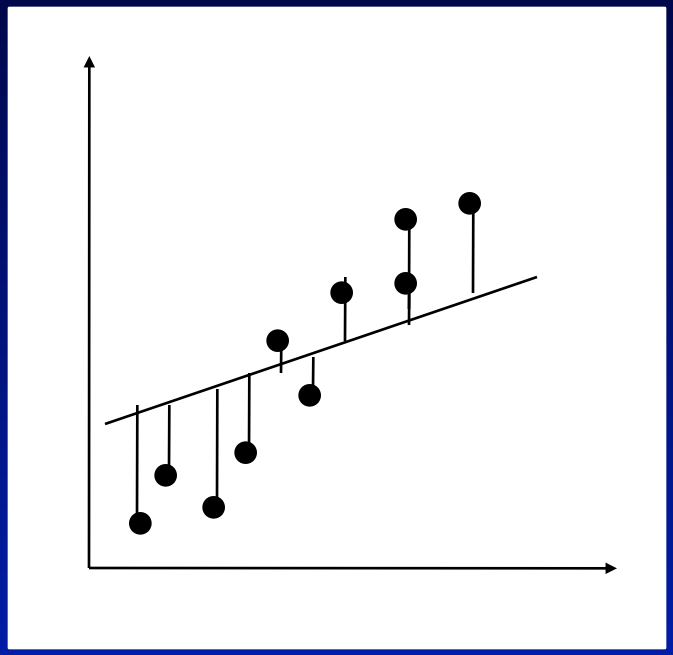
Modeling: An example



$$y = ax + b$$

Predictions based on model

Model Fit, parameter estimation

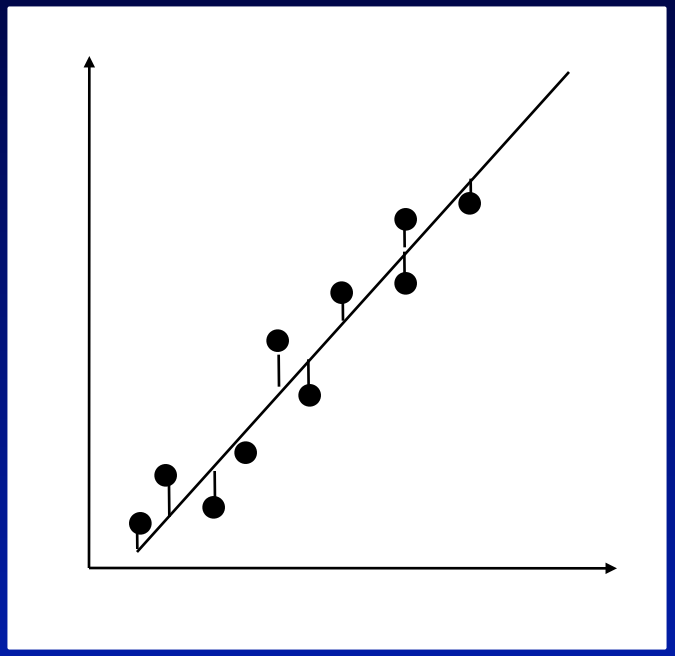


$$y = ax + b$$

Measure of how well the model fits the data: sum of squared errors (SSE)

Best parameter estimates: those that give the smallest SSE (least squares model fitting)

Model Fit , parameter estimation



$y = 1.24x - 0.56$

Measure of how well the model fits the data: sum of squared errors (SSE)

Best parameter estimates: those that give the smallest SSE (least squares model fitting)

The maximum likelihood approach I

- Starting point:

  You have some observed data and a probabilistic model for how the observed data was produced

  Having a probabilistic model of a process means you are able to compute the probability of any possible outcome (given a set of specific values for the model parameters).

- Example:
  - Data: result of tossing coin 10 times - 7 heads, 3 tails
  - Model: coin has probability p for heads, 1-p for tails.

    The probability of observing h heads among n tosses is:

$$P(\text{h heads}) = \binom{n}{h} p^h (1-p)^{n-h}$$

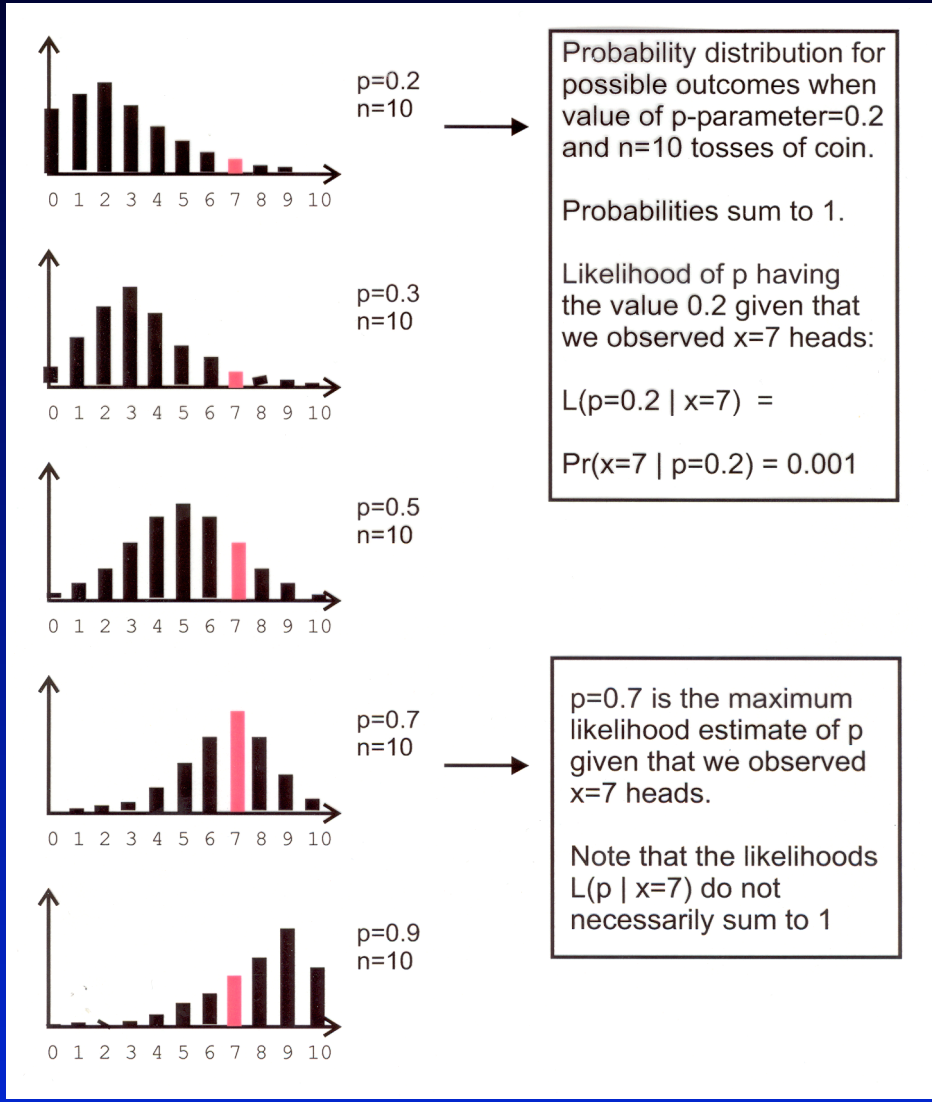
- Goal:

  You want to find the best estimate of the (unknown) parameter values based on the observations. (here the only parameter is “p”)

The maximum likelihood approach II

- Likelihood (Model) = Probability (Data | Model)
- Maximum likelihood:
  Best estimate is the set of parameter values which gives the highest possible likelihood.

Maximum likelihood: coin tossing example



Data: result of tossing coin 10 times - 7 heads, 3 tails

Model: coin has probability p for heads, 1-p for tails.

$$P(\text{data}) = \binom{10}{7} p^7 (1-p)^3$$

Probabilistic modeling applied to phylogeny

- Observed data: multiple alignment of sequences

```
H.sapiens globin          A G G G A T T C A
M.musculus globin         A C G G T T T - A
R.rattus globin           A C G G A T T - A
```

- Probabilistic model:
  - A model of (hypothesis about) how one ancestral sequence has evolved into the three sequences that are present in the alignment

- Probabilistic model parameters (simplest case):
  - Tree topology and branch lengths
  - Nucleotide frequencies: π_A, π_C, π_G, π_T
  - Nucleotide-nucleotide substitution rates (or substitution probabilities):

$$Q = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & -3\alpha & \alpha & \alpha & \alpha \\ C & \alpha & -3\alpha & \alpha & \alpha \\ G & \alpha & \alpha & -3\alpha & \alpha \\ T & \alpha & \alpha & \alpha & -3\alpha \end{array}$$

Rate matrix

$$\Rightarrow P(t) = e^{Qt} = \begin{bmatrix} P_{AA} & P_{AC} & P_{AG} & P_{AT} \\ P_{CA} & P_{CC} & P_{CG} & P_{CT} \\ P_{GA} & P_{GC} & P_{GG} & P_{GT} \\ P_{TA} & P_{TC} & P_{TG} & P_{TT} \end{bmatrix}$$

Probability matrix (function of time t)

Other models of nucleotide substitution

TABLE 3.1 Models of nucleotide substitution

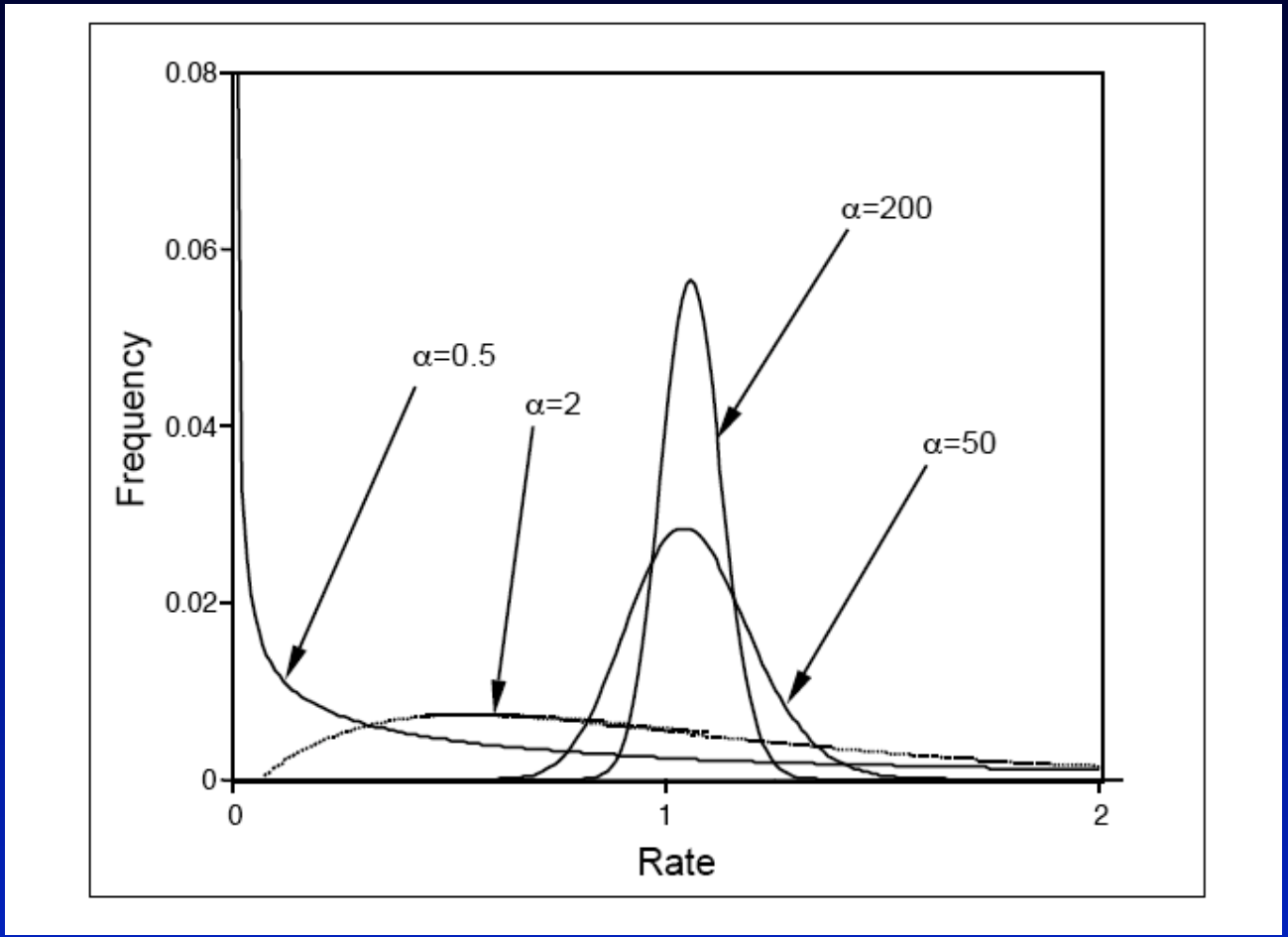
O\S[a]	A	T	C	G
a. Two-parameter model (Kimura 1980)				
A	$1-\alpha-2\beta$	β	β	α
T	β	$1-\alpha-2\beta$	α	β
C	β	α	$1-\alpha-2\beta$	β
G	α	β	β	$1-\alpha-2\beta$
b. Four-parameter model (Blaisdell 1985)				
A	$1-\alpha-2\gamma$	γ	γ	α
T	δ	$1-\alpha-2\delta$	α	δ
C	δ	β	$1-\beta-2\delta$	δ
G	β	γ	γ	$1-\beta-2\gamma$
c. Six-parameter model (Kimura 1981a)				
A	$1-2\alpha-\gamma$	γ	α	α
T	δ	$1-2\alpha-\delta$	α	α
C	β	β	$1-2\beta-\varepsilon$	ε
G	β	β	ξ	$1-2\beta-\xi$
d. Nine-parameter model				
A	$1-g_T\beta_1-g_C\gamma_1-g_G\alpha_1$	$g_T\beta_1$	$g_C\gamma_1$	$g_G\alpha_1$
T	$g_A\beta_1$	$1-g_A\beta_1-g_C\alpha_2-g_G\gamma_2$	$g_C\alpha_2$	$g_G\gamma_2$
C	$g_A\gamma_1$	$g_T\alpha_2$	$1-g_A\gamma_1-g_T\alpha_2-g_G\beta_2$	$g_G\beta_2$
G	$g_A\alpha_1$	$g_T\gamma_2$	$g_C\beta_2$	$1-g_A\alpha_1-g_T\gamma_2-g_C\beta_2$
e. General model				
A	$1-\alpha_{12}-\alpha_{13}-\alpha_{14}$	α_{12}	α_{13}	α_{14}
T	α_{21}	$1-\alpha_{21}-\alpha_{23}-\alpha_{24}$	α_{23}	α_{24}
C	α_{31}	α_{32}	$1-\alpha_{31}-\alpha_{32}-\alpha_{34}$	α_{34}
G	α_{41}	α_{42}	α_{43}	$1-\alpha_{41}-\alpha_{42}-\alpha_{43}$

[a]O, Original nucleotide; S, substitute nucleotide.

More elaborate models of evolution

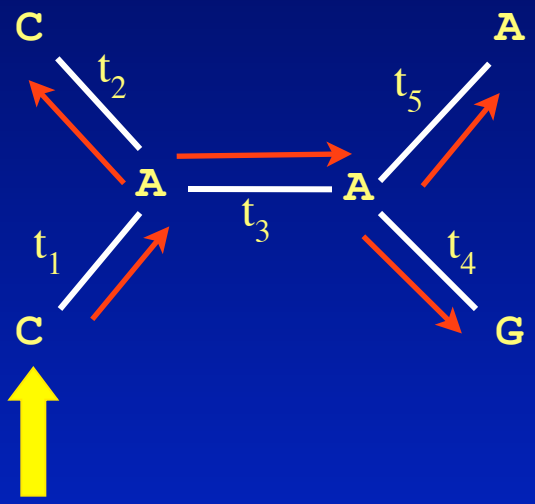
- Codon-codon substitution rates
  (64 x 64 matrix of codon substitution rates)
- Different mutation rates at different sites in the gene
  (the “gamma-distribution” of mutation rates)
- Molecular clocks
  (same mutation rate on all branches of the tree).
- Different substitution matrices on different branches of the tree
  (e.g., strong selection on branch leading to specific group of animals)
- Etc., etc.

Different rates at different sites: the gamma distribution



Computing the probability of one column in an alignment given tree topology and other parameters


```
A C G G A T T C A
A C G G T T T - A
A A G G A T T - A
A G G G T T T - A
```

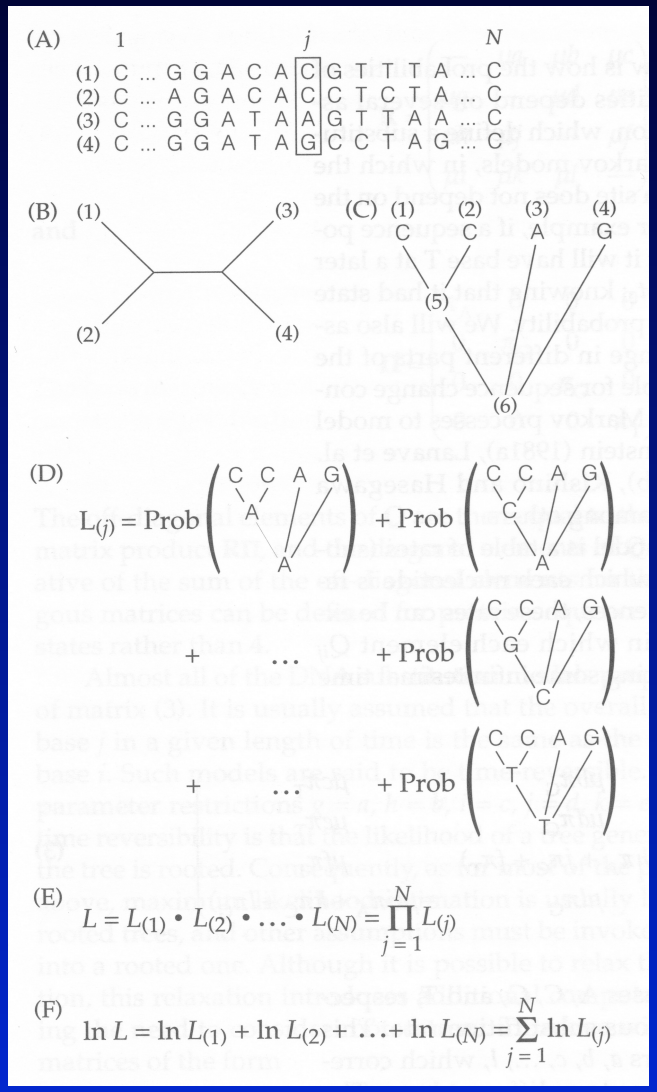


Columns in alignment contain homologous nucleotides

Assume tree topology, branch lengths, and other parameters are given. For now, assume ancestral states were A and A (we'll get to the full computation on next slide). Start computation at any internal or external node. Arrows indicate "direction" of computations ("flowing" away from the starting point).

$$Pr = \pi_C \, P_{CA}(t_1) \, P_{AC}(t_2) \, P_{AA}(t_3) \, P_{AG}(t_4) \, P_{AA}(t_5)$$

Computing the probability of an entire alignment given tree topology and other parameters



- Probability must be summed over all possible combinations of ancestral nucleotides.

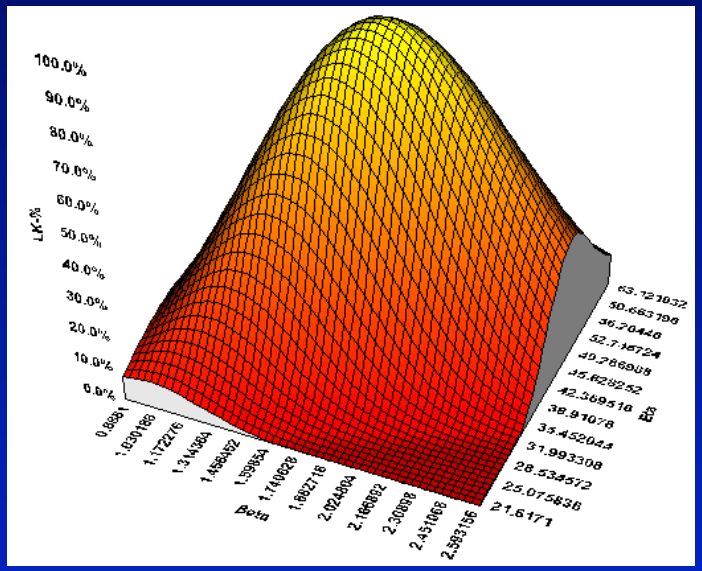
(Here we have two internal nodes giving 16 possible combinations)

- Probability of individual columns are multiplied to give the overall probability of the alignment, i.e., the likelihood of the model.

- In phylogeny software these computations are done using summation of the logs of the probabilities ("log likelihoods"), because multiplication of the large number of probability terms may lead to underflow (computer problems caused by very small numbers).

Maximum likelihood phylogeny

- Data:
  - sequence alignment
- Model parameters:
  - nucleotide frequencies, nucleotide substitution rates, tree topology, branch lengths.



- Choose random initial values for all parameters, compute likelihood
- Change parameter values slightly in a direction so likelihood improves
- Repeat until maximum found
- Results:
  - (1) ML estimate of tree topology
  - (2) ML estimate of branch lengths
  - (3) ML estimate of other model parameters
  - (4) Measure of how well model fits data (likelihood).

Tree 1			
Node 1	Node 2	Likelihood	Sum
A	A	0.0008091	0.0025323
A	C	0.0008356	
A	G	0.0008611	
A	T	0.0000265	
C	A	0.0000003	0.0000627
C	C	0.0000566	
C	G	0.0000056	
C	T	0.0000002	
G	A	0.0000005	0.0000680
G	C	0.0000072	
G	G	0.0000601	
G	T	0.0000002	
T	A	0.0000001	0.0000055
T	C	0.0000018	
T	G	0.0000019	
T	T	0.0000017	
Sum		0.0026683	

Ancestral reconstruction:

Node1 = A

Ancestral reconstruction: experimental analysis of extinct molecules

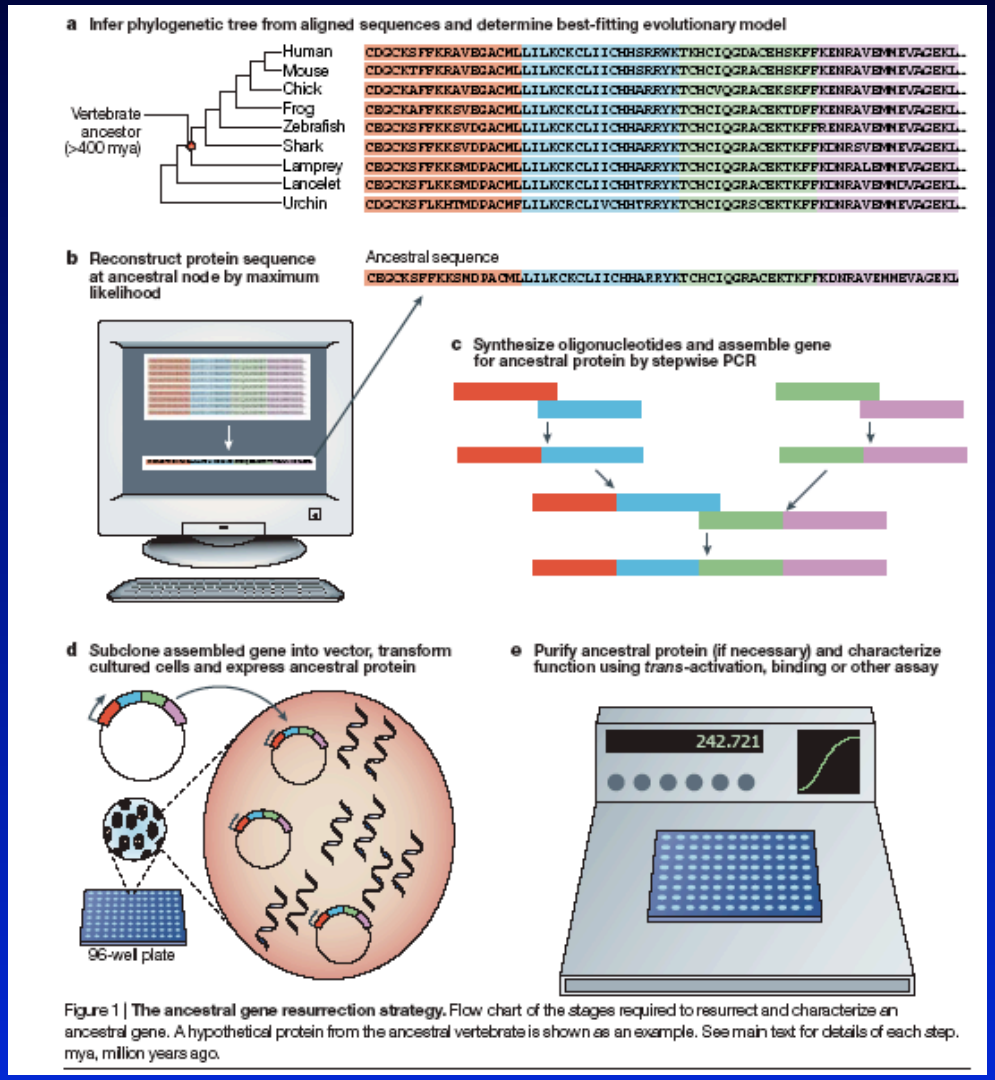


Figure 1 | **The ancestral gene resurrection strategy.** Flow chart of the stages required to resurrect and characterize an ancestral gene. A hypothetical protein from the ancestral vertebrate is shown as an example. See main text for details of each step. mya, million years ago.

Ancestral reconstruction

Table 1 | **Ancestral genes resurrected***

Extant genes	Ancestral gene resurrected	Approximate age of ancestor (years)	Inference method	Refs
Digestive ribonucleases	Ancestral orthologue in LCA of buffalo and ox	5–10 million	Parsimony	38
L1 retroposons in mouse	Ancestral paralogue‡ in mouse genome	"several million"	Consensus	7
Digestive ribonucleases	Ancestral orthologue in LCA of artiodactyls	~40 million	Parsimony	22
Chymase proteases	Ancestral orthologue in LCA of mammals	~80 million	Parsimony	43
Tc1/mariner transposons	Ancestral paralogue in genomes of 8 salmonids	~10 million	Consensus	6
Immune RNases	Ancestral orthologue in LCA of 'higher primates'	31 million	Parsimony, Bayesian distance	27
Pax§ transcription factors	Ancestral paralogue (older than the protostome‖–deuterostome¶ ancestor)	600–1,000 million	Bayesian distance	26
Vertebrate rhodopsins	Ancestral orthologue in LCA of archosaurs#	240 million	Maximum likelihood	17
Vertebrate short-wave rhodopsins	Ancestral orthologue in LCA of bony vertebrates	>400 million	Maximum likelihood	44
Steroid hormone receptors	Ancestral paralogue (older than the protostome–deuterostome ancestor)	600–1,000 million	Maximum likelihood	18
Elongation factor EF-Tu	Ancestral orthologue in LCA of eubacteria	>1 billion	Maximum likelihood	20

*Papers that have inferred ancestral sequences and synthesized them for functional analysis. (Studies that used directed mutagenesis to examine the effects of isolated replacements are not included.) ‡Paralogue, evolutionarily related genes that are produced by gene duplication. §Pax, paired box protein-encoding gene. ‖Protostome, a bilaterian animal, the mouth of which develops before the anus during embryogenesis. Protostomes include arthropods, molluscs and worms. ¶Deuterostome, a bilaterian animal, the mouth of which forms after the anus during embryogenesis. Deuterostomes include chordates, hemichordates and echinoderms. #Archosaur, a member of the animal taxon that includes all crocodiles, birds and extinct dinosaurs. LCA, last common ancestor.

Ancestral reconstruction: dinosaur night vision

Despite its great age, the ancestral rhodopsin functioned well, carrying out all the individual steps that are required for visual function in dim light as effectively as the extant proteins in mammals, which generally have good night vision. Specifically, the ancestral protein bound the visual chromophore 11-*cis*-retinal and, when exposed to light, activated the G-protein transducin at a rate similar to that of bovine rhodopsin. These results are consistent with the hypothesis that the ancestral archosaur possessed the ability — at the molecular level at least — to see well in dim light, and might have been active at night. This insight, of course, could never have been drawn from fossils or any other non-molecular evidence about the behaviour of ancient dinosaurs.

Ancestral reconstruction: thermostability of ancestral proteins

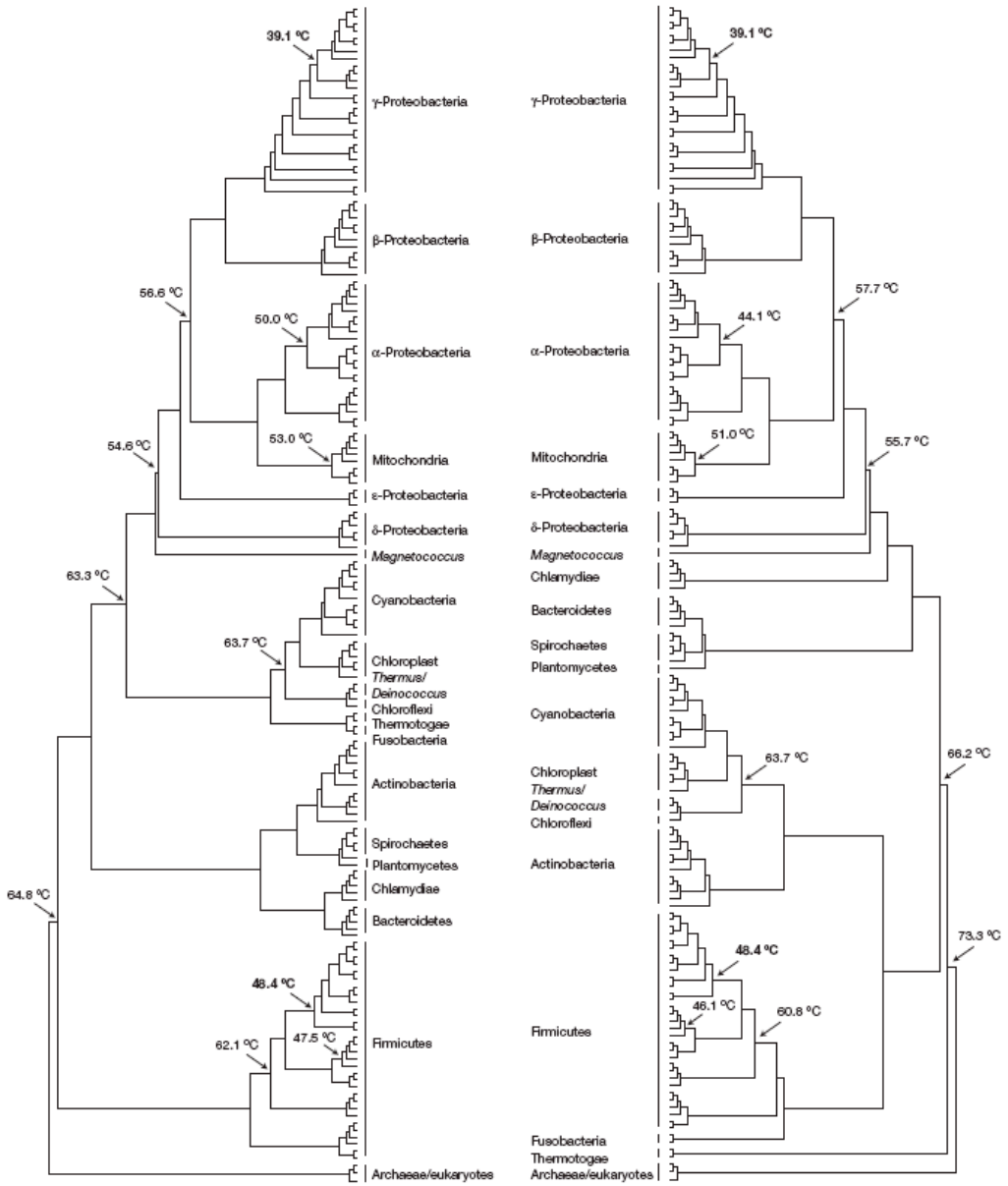


Figure 1 | Cladograms of the trees used to resurrect EF proteins. The phylogeny on the left follows ref. 17, whereas that on the right follows ref. 16. Melting temperatures for ancestral EF proteins are shown at their corresponding nodes. Temperatures in bold represent identical sequences at analogous nodes between the two trees. Errors associated with measurements of circular dichroism are negligible (see Supplementary Information).

Ancestral reconstruction: thermostability of ancestral proteins

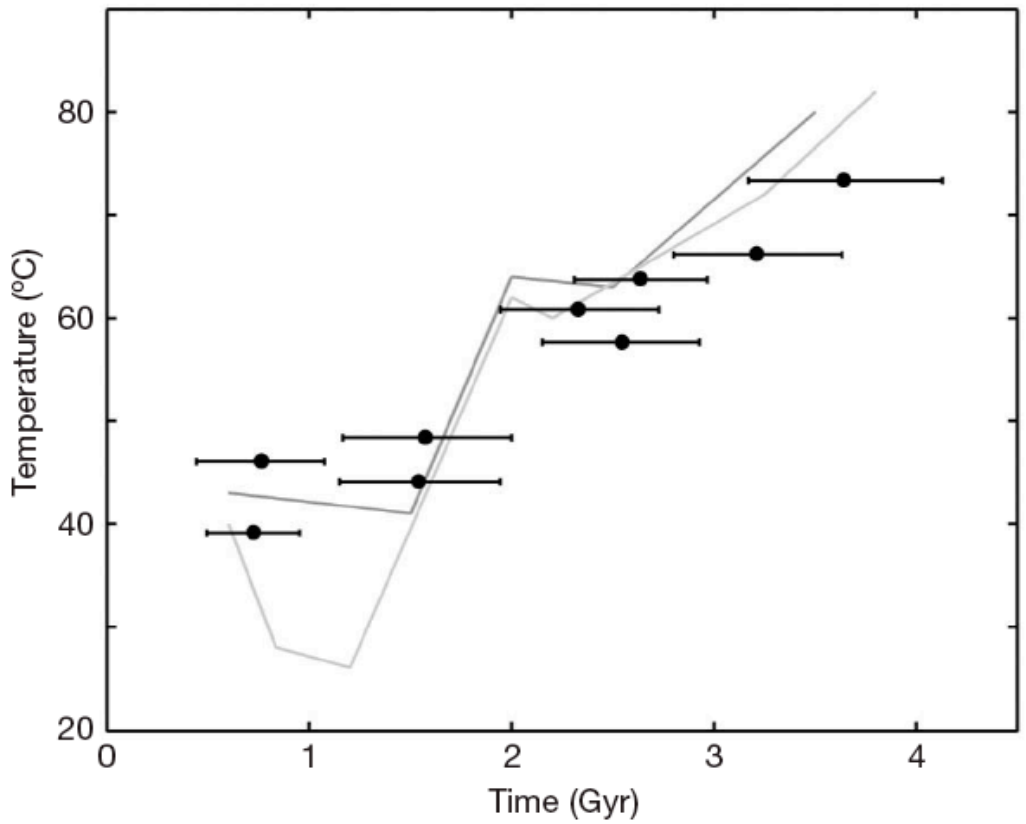


Figure 3 | Plot of ancestral EF melting temperatures against geological time. Molecular clock estimates are shown with their confidence intervals (horizontal bars) from ref. 16, using a 2.3-Ga minimum constraint for the Great Oxidation Event. Solid lines are temperature curves of the ancient ocean inferred from maximum $\delta^{18}O$ (light grey[3,4], dark grey[5]). Although not shown, an analogous trend is seen with $\delta^{30}Si$ isotopes[5].

Phylogeny and ancestral reconstruction for manuscripts

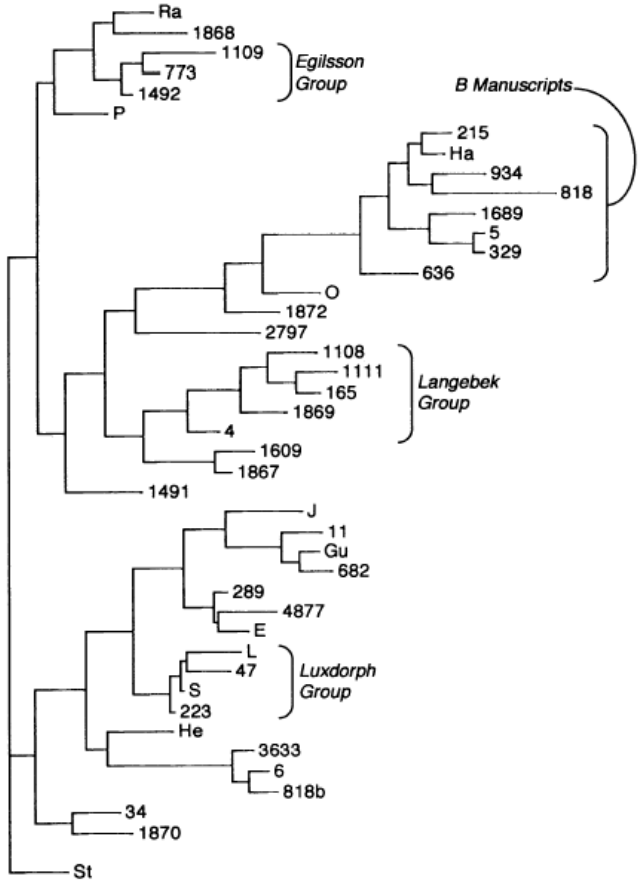
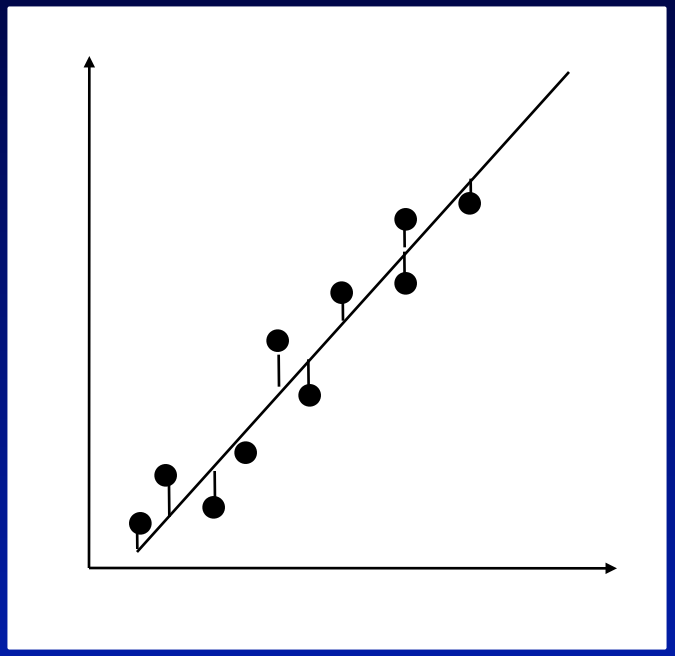


Fig. 7 - A stemma of the Old Norse narrative *Svipdagsmál,* from Robinson & O'Hara (1993, in press). This stemma was produced with the cladistic analysis software PAUP (Swofford, 1991), and it is very similar to the stemma produced by Robinson alone using traditional non-cladistic means (Robinson, 1991). This tree was generated much more quickly, however, thereby allowing the textual scholar more time for critical study and analysis of the result.

Model Selection?

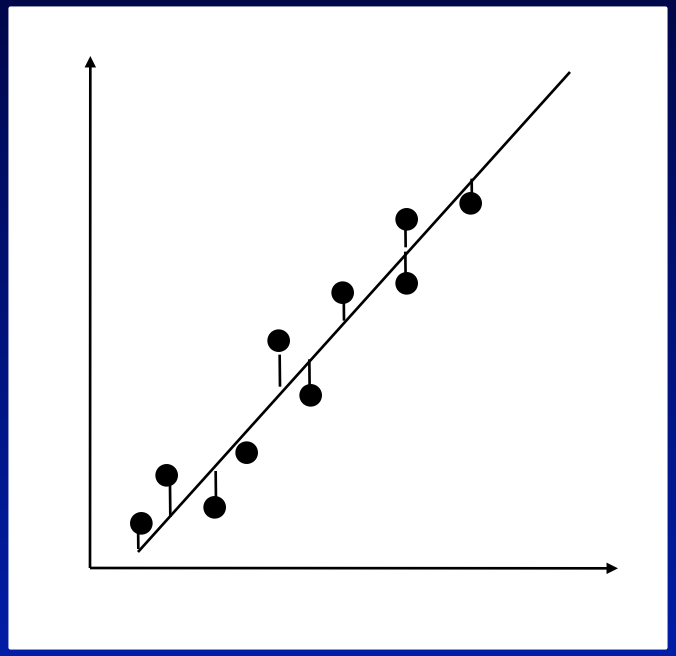


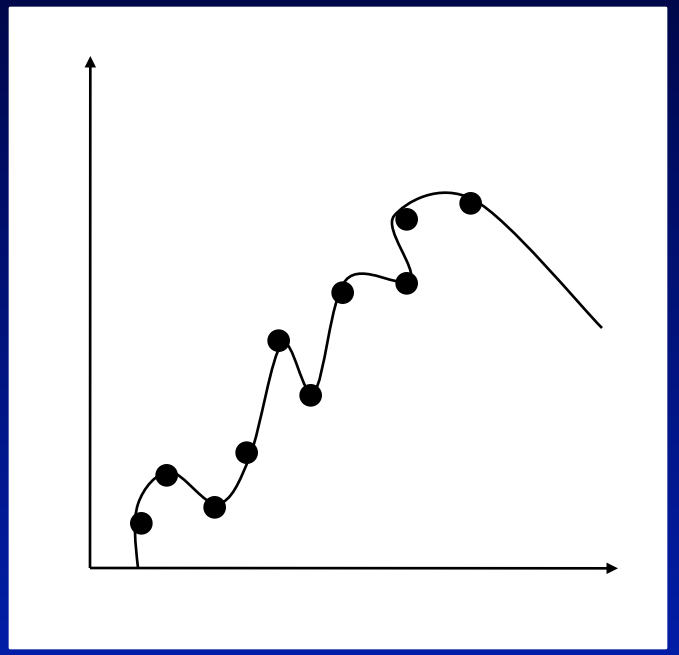
$y = 1.24x - 0.56$

Measure of fit between model and data (e.g., SSE, likelihood, etc.)

How do we compare different *types* of models?

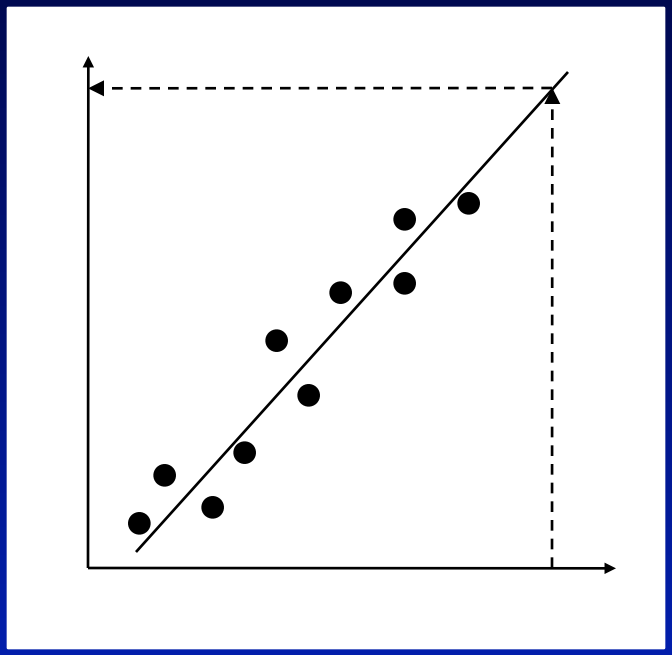
Model Selection:
How Do We Choose Between Different Types of Models?





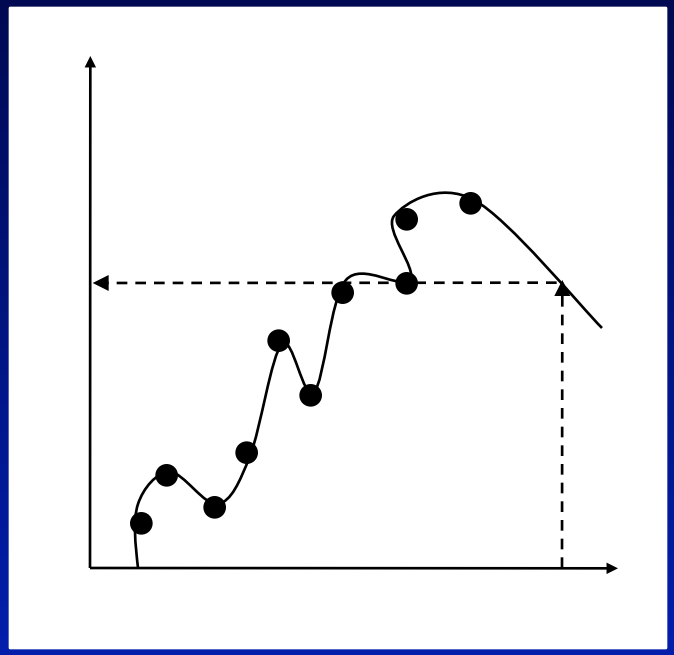
Select model with best fit?

Over-fitting: For nested models, more parameters always result in a better fit to the data, but not necessarily in a better description



$y = ax + b$

2 parameter model
Good description, poor fit



$y = ax^6+bx^5+cx^4+dx^3+ex^2+fx+g$

7 parameter model
Poor description, good fit

Selecting the best model: the likelihood ratio test

- The fit of two alternative models can be compared using the ratio of their likelihoods:

$$LR = \frac{P(Data|M1)}{P(Data|M2)} = \frac{L,M1}{L,M2}$$

- Note that LR > 1 if model 1 has the highest likelihood

- For nested models it can be shown that if the simplest ("null") model is true, then

$$\Delta = \ln(LR^2) = 2*\ln(LR) = 2* (\ln L,M1 - \ln L,M2)$$

follows a X^2 distribution with degrees of freedom equal to the number of extra parameters in the most complicated model.

This makes it possible to perform stringent statistical tests to determine which model (hypothesis) best describes the data

Asking biological questions in a likelihood ratio testing framework

- Fit two alternative, nested models to the data.
- Record optimized likelihood and number of free parameters for each fitted model.
- Test if alternative (parameter-rich) model is *significantly* better than null-model (i.e., the simplest model), given number of additional parameters (n_{extra}):

1. Compute $\Delta = 2 \times (\ln L_{Alternative} - \ln L_{Null})$
2. Compare Δ to X^2 distribution with n_{extra} degrees of freedom

- Depending on models compared, different biological questions can be addressed (presence of molecular clock, presence of positive selection, difference in mutation rates among sites or branches, etc.)

Model Selection Using the Akaike Information Criterion (AIC)

- Fit a set of alternative models to the data.
- Record maximized log likelihood (lnL) and number of free parameters (K) for each fitted model.
- For each model compute AIC according to this formula:

$$AIC = -2 \times \ln L + 2 \times K$$

- Models can now be ranked according to AIC: Lower AIC is better.

Model	ℓ	K	AIC_c
TN93+I+Γ	5441.4600	78	11045.5888
TIM+I+Γ	5441.3765	79	11047.5965
HKY85+I+Γ	5443.6729	77	11047.8422
K81uf+I+Γ	5443.5566	78	11049.7821
GTR+I+Γ	5440.9150	81	11051.0301
TVM+I+Γ	5442.7393	80	11052.4991
TN93+Γ	5448.6792	77	11057.8549
HKY85+Γ	5450.5068	76	11059.3402
TIM+Γ	5448.6577	78	11059.9843
K81uf+Γ	5450.4883	77	11061.4730
GTR+Γ	5448.0298	80	11063.0802
TVM+Γ	5449.6685	79	11064.1804
TN93+I	5470.7568	77	11102.0102
TIM+I	5470.7417	78	11104.1522
GTR+I	5470.3452	80	11107.7110

AIC: basis in information theory

- AIC is firmly based on information theory.
- Briefly, it is an estimate of the expected, relative Kullback-Leibler distance between the true model and the approximating model.

$$K\text{-}L = I(f,g) = \int f(x)\log\left(\frac{f(x)}{g(x \mid \theta)}\right)dx$$

Full reality Approximating model

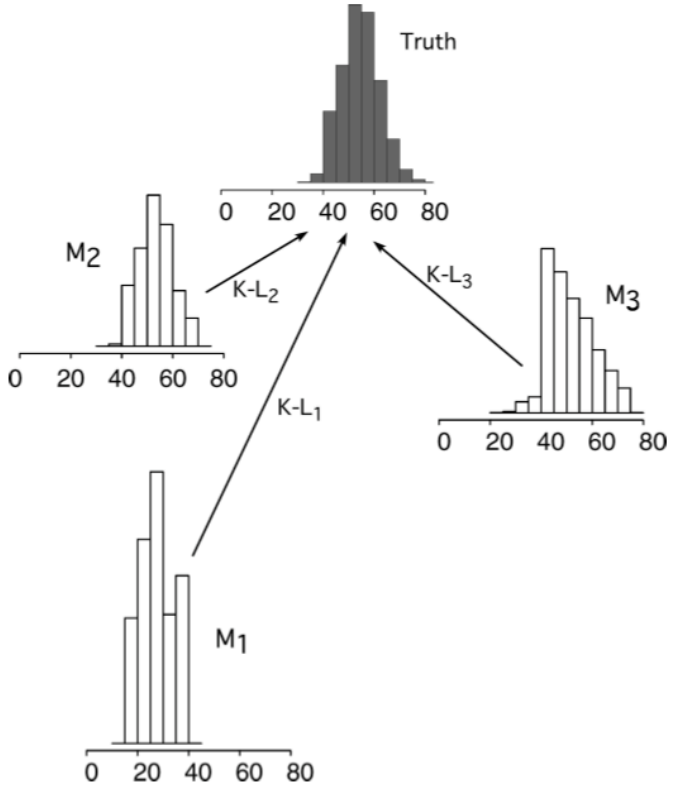


FIGURE 4. The Kullback-Leibler distance. The K-L distance aims to represent how close a model is to the truth. Here, M_2 is the candidate model that best approximates truth and therefore it is the model with the smallest K-L distance. The AIC chooses the candidate model with the smallest *expected* K-L distance.

Model Selection Using the AIC: computation of model probabilities

- From the relative AIC values it is furthermore possible to compute so-called Akaike weights:

$$\Delta\mathrm{AIC}_i = \mathrm{AIC}_i - \min \mathrm{AIC}$$

$$w_i = \frac{\exp(-1/2\Delta_i)}{\sum_{r=1}^{R} \exp(-1/2\Delta_r)}$$

Model	ℓ	K	AIC_c	$\Delta\mathrm{AIC}_c$	w
TN93+I+Γ	5441.4600	78	11045.5888	0.0000	0.5221
TIM+I+Γ	5441.3765	79	11047.5965	2.0077	0.1913
HKY85+I+Γ	5443.6729	77	11047.8422	2.2534	0.1692
K81uf+I+Γ	5443.5566	78	11049.7821	4.1934	0.0641
GTR+I+Γ	5440.9150	81	11051.0301	5.4413	0.0344
TVM+I+Γ	5442.7393	80	11052.4991	6.9103	0.0165
TN93+Γ	5448.6792	77	11057.8549	12.2661	0.0011
HKY85+Γ	5450.5068	76	11059.3402	13.7514	0.0005
TIM+Γ	5448.6577	78	11059.9843	14.3955	0.0004
K81uf+Γ	5450.4883	77	11061.4730	15.8843	0.0002
GTR+Γ	5448.0298	80	11063.0802	17.4914	0.0001
TVM+Γ	5449.6685	79	11064.1804	18.5917	0.0000
TN93+I	5470.7568	77	11102.0102	56.4214	0.0000
TIM+I	5470.7417	78	11104.1522	58.5635	0.0000
GTR+I	5470.3452	80	11107.7110	62.1223	0.0000

- Akaike weight can be interpreted as the conditional probability that a model is the K-L best one, given the data and the initial set of models.

Probabilities as extended logic

- Polya, Cox, Jeffreys, Jaynes: probabilities are the only consistent basis for plausible reasoning (reasoning when there is insufficient information for deductive reasoning).
- Probabilities should form basis of all scientific inference
- Difference between probability interpretations:
  - "Frequentist": probability is long-run frequency of event in repeatable experiment
  - "Bayesian": probability is way of quantifying uncertainty
- Attaching probabilities to models allow us to perform multimodel inference and model averaging

Model selection as a general strategy for answering scientific questions

- Construct comprehensive set of plausible alternative hypotheses for how the system under investigation works (but not too many)
- Phrase the hypotheses as mathematical models
- Assess evidence for all hypotheses by computing model probabilities
- Make conclusions, predictions, etc based on model probabilities
- Very different from null hypothesis testing approach (where you assess fit of single, implausible model that you don't believe is true...)

Multimodel Inference: Basing Conclusions on More Than Just the Best Model

- Prediction: More robust predictions can be made by taking a weighted average of the predictions made by all models. (Weight = model probability)
- Model-averaging: More reliable estimates of parameter values can be obtained by taking a weighted average over the sub-set of fitted models that contain the parameter.
- Relative importance of parameters: the importance of a parameter can be estimated by summing the probabilities of those models that contain it.

AIC example:
Which model fits best: JC or K2P?

	A	C	G	T
A	-	α	α	α
C	α	-	α	α
G	α	α	-	α
T	α	α	α	-

Jukes and Cantor model (JC):

All nucleotides have same frequency
All substitutions have same rate
K = 1 parameter

	A	C	G	T
A	-	β	α	β
C	β	-	β	α
G	α	β	-	β
T	β	α	β	-

Kimura 2 parameter model (K2P):

All nucleotides have same frequency
Transitions and transversions have different rate
K = 2 parameters

Note: in principle each branch length in the tree also has an associated free parameter, but we ignore these here since they cancel out (the tree is the same in the two cases)

Note 2: depending on how you phrase the problem, JC and K2P can be said to have K=0 and K=1

Likelihood ratio test example: Which model fits best: JC or K2P?

Starting point: set of DNA sequences, fit JC and K2P models to data, record likelihoods

JC: lnL = -2034.3, K = 1
K2P: lnL = -2026.2, K = 2

Assess evidence by computing model probabilities:

(1) Compute AIC = -2 lnL + 2K:

JC: AIC = -2 x -2034.3 + 2 x 1 = 4070.6
K2P: AIC = -2 x -2026.2 + 2 x 2 = 4056.4 <= Best model (smallest AIC)

(2) Compute $\Delta \mathrm{AIC}_i = \mathrm{AIC}_i - \min \mathrm{AIC}$

JC: 4070.6 - 4056.4 = 14.2
K2P: 4056.4 - 4056.4 = 0

Likelihood ratio test example: Which model fits best: JC or K2P?

(3) Compute model probabilities:

$$w_i = \frac{\exp(-1/2\Delta_i)}{\sum_{r=1}^{R} \exp(-1/2\Delta_r)}$$

JC: numerator = exp(-0.5 x 14.2) = 0.000825
K2P: numerator = exp(-0.5 x 0) = 1

Sum (denominator) = 1 + 0.000825 = 1.000825

=>

P(JC) = 0.000825 / 1.000825 = 0.0008 (0.08 %)
P(K2P) = 1 / 1.000825 = 0.9992 (99.92 %) <= Strongly supported (about 1250 x stronger)

Bayesian model comparison

- We can compare hypotheses, very much like we can compare possible parameter values within a single hypothesis:

$$P(H_1|D) = \frac{P(D|H_1)P(H_1)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)}$$

- Let's assign equal prior probabilities to the two hypotheses: $P(H_1) = P(H_2) = 0.5$

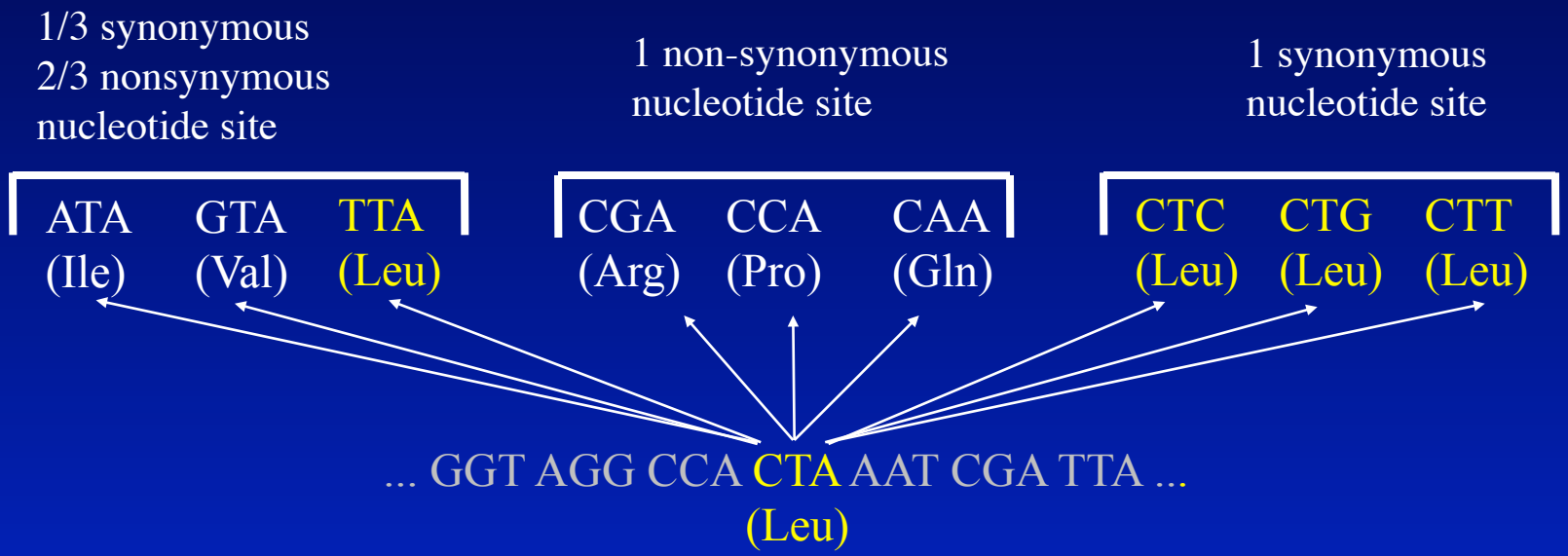
- The likelihoods $P(D|H_1)$ and $P(D|H_2)$ are in fact terms we have computed before: They are the denominators from the posteriors for H_1 and H_2 respectively (i.e, they are each a sum of 11 prior x likelihood terms):

$$P(p_+ = 0.4|D) = \frac{P(D|p_+ = 0.4)P(p_+ = 0.4)}{\sum_{p=0.0}^{1.0} P(D|p_+ = p)P(p_+ = p)}$$

$P(D|H_1)$

Positive selection I: synonymous and non-synonymous mutations

- 20 amino acids, 61 codons
- Most amino acids encoded by more than one codon
  - Not all mutations lead to a change of the encoded amino acid
  - ”Synonymous mutations” are rarely selected against



Positive selection II: non-synonymous and synonymous mutation rates contain information about selective pressure

- dN: rate of non-synonymous mutations per non-synonymous site
- dS: rate of synonymous mutations per synonymous site

- Recall: Evolution is a two-step process:
  - (1) Mutation (random)
  - (2) Selection (non-random)

- Randomly occurring mutations will lead to dN/dS=1.
- Deviations from this most likely caused by subsequent selection.

- dN/dS < 1: Higher rate of synonymous mutations: negative (purifying) selection
- dN/dS > 1: Higher rate of non-synonymous mutations: positive selection

Today's exercise: positive selection in HIV?

- Fit two alternative models to HIV data:
  - M1: two classes of codons with different dN/dS ratios: dN/dS<1 ⬡ dN/dS=1 ⬡
  - M2: three distinct classes with different dN/dS ratios: dN/dS<1 ⬡ dN/dS=1 ⬡ dN/dS>1 ⬡
- Compute model probabilities to assess the evidence for M2 versus M1
- If M2 much better than M1 then you have statistical evidence for positive selection.
- Most likely reason: immune escape (i.e., sites must be in epitopes)

⬡ : Codons showing dN/dS > 1: likely epitopes